BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 5 2009, pages 684–686
doi:10.1093/bioinformatics/btp026

Systems biology

KEA: kinase enrichment analysis

Alexander Lachmann and Avi Ma'ayan*

Department of Pharmacology and Systems Therapeutics, Systems Biology Center in New York, Icahn Medical Institute, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY 10029, USA

Received on November 8, 2008; revised on December 19, 2008; accepted on January 8, 2009

Advance Access publication January 28, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Multivariate experiments applied to mammalian cells often produce lists of proteins/genes altered under treatment versus control conditions. Such lists can be projected onto prior knowledge of kinase–substrate interactions to infer the list of kinases associated with a specific protein list. By computing how the proportion of kinases, associated with a specific list of proteins/genes, deviates from an expected distribution, we can rank kinases and kinase families based on the likelihood that these kinases are functionally associated with regulating the cell under specific experimental conditions. Such analysis can assist in producing hypotheses that can explain how the kinome is involved in the maintenance of different cellular states and can be manipulated to modulate cells towards a desired phenotype.

Summary: Kinase enrichment analysis (KEA) is a web-based tool with an underlying database providing users with the ability to link lists of mammalian proteins/genes with the kinases that phosphorylate them. The system draws from several available kinase–substrate databases to compute kinase enrichment probability based on the distribution of kinase–substrate proportions in the background kinase–substrate database compared with kinases found to be associated with an input list of genes/proteins.

Availability: The KEA system is freely available at http://amp.pharm.mssm.edu/lib/kea.jsp

Contact: avi.maayan@mssm.edu

1 INTRODUCTION

Protein phosphorylation causes the addition of a phosphate group onto serine, threonine or tyrosine amino-acid residues of proteins. Phosphorylations are precise reversible changes that are used to regulate intracellular events such as protein complex formation, cell signaling, cytoskeleton remodeling and cell cycle control. Consequently, protein kinases, which are responsible for the phosphorylations, play an important role in controlling protein function, cellular machine regulation and information transfer through cell signaling pathways. Kinase activities therefore have definitive regulatory effects on a broad variety of biological processes, in which activated kinases typically target a large number of different substrate proteins. There are over 500 protein kinases encoded in the human genome, and it is approximated that 40% of all proteins are phosphorylated at some stage in different cell types and at different cell states (Manning *et al*., 2002). Furthermore, kinases regulate each other through phosphorylation, resulting in a complex web of regulatory relations (Ma'ayan *et al*., 2005).

High-throughput techniques such as stable isotope labeling coupled with affinity purification and mass-spectrometry proteomics are now able to identify phosphorylation sites on multiple proteins under different experimental conditions. Databases that integrate the results from such studies are emerging, e.g. phosphosite (Hornbeck *et al*., 2004). However, such data does not provide the kinases responsible for the phosphorylation. Several resources are available to link identified phosphorylation sites to the kinases that are most likely responsible for protein phosphorylations (Huang *et al*., 2005; Linding *et al*., 2008). For example, NetworKIN (Linding *et al*., 2007; Linding *et al*., 2008) uses an algorithm to predict the most probable kinase that is responsible for phosphorylating an identified phosphosite. The NetworKIN algorithm is accompanied with a database containing ~1450 predicted mammalian substrates that are mapped to 73 upstream protein kinases belonging to 21 kinase families. Although useful, the coverage of this dataset is not comprehensive enough for kinase statistical enrichment analysis. To achieve more comprehensive prior knowledge kinase–substrate dataset, large enough for statistical enrichment analysis, we merged interactions from several other online sources reporting mammalian kinase–substrate relations. Additionally, we included binary protein—-protein interactions involving kinases from protein–protein interaction databases as these were recently proposed to be highly enriched in kinase–substrate relations: in a recent study that identified ~14 000 phosphosites at different stages of the cell cycle in Hela cells (Dephoure *et al*., 2008) it was shown that many phosphosites experimentally identified using phosphoproteomics can be associated with four known kinases (CDC2, PLK1, Aurora-B and Aurora-A) using the literature-based protein–protein interactions from the HPRD database (Mishra *et al*., 2008). Hence, having a large background knowledge dataset of kinase–substrate interactions and protein–protein interactions that involve kinases, we can associate large lists of proteins/genes with many kinases that phosphorylate them. This allows the computation of statistical enrichment which can be used to suggest the kinases that are most likely to be involved in regulating the proteins/genes from a list generated under specific experimental conditions.

2 IMPLEMENTATION

We first constructed a database that consolidates kinase–substrate interactions from multiple online sources. We integrated data describing kinase–substrate interactions from NetworKIN

*To whom correspondence should be addressed.

(Linding *et al*., 2008), Phospho.ELM (Diella *et al*., 2004), MINT (Chatr-aryamontri *et al*., 2007), HPRD (Mishra *et al*., 2008), PhosphoPoint (Yang *et al*., 2008) and Swiss-Prot (Quintaje and Orchard, 2008) as well as phosphorylation interactions we manually previously extracted from literature (Ma'ayan *et al*., 2005). The NetworKIN database contains 3847 kinase–substrate unique pairs made of 73 kinases (21 families) linked to 1452 substrates. HPRD contains 1794 kinase–substrate pairs made of 229 kinases linked to 864 substrates. Phospho.Elm has 1451 interactions between 225 kinases and 784 substrates. MINT has 269 interactions between 145 kinases and 184 substrates. In phosphoPoint there are 436 kinases, 3076 substrates, 9251 kinase–substrate relations from which only 1587 are unique in this dataset, while the rest overlaps with the other databases. In Ma'ayan *et al*., there are 66 interactions between 19 kinases and 43 substrates. There is some overlap among these sources such that the number of unique kinase–substrate relations totals 6414 links between 352 kinases and 2014 substrates in the combined dataset. We consolidated interactions from mouse and rat into human by converting all protein/gene IDs to human Entrez gene symbols. Each kinase–substrate data record is associated with a specific kinase, kinase family and kinase subfamily. To group kinases into families, we used the kinome tree from Manning *et al*. (2002) where kinases are classified into 10 major classes and 119 families. To further increase the size of our background dataset, we included all direct protein–protein interactions involving kinases from HPRD (Mishra *et al*., 2008) and MINT (Chatr-aryamontri *et al*., 2007). By this expansion the current dataset contains a total of 11 923 interactions between 445 kinases having 3995 substrates.

The analysis begins with an input list of gene symbols entered by the user for kinase enrichment analysis (KEA). Before performing the KEA, we remove all input entries that do not match a substrate in the consolidated background kinase–substrate dataset. This step is necessary for achieving proportional comparison. The expected value for a randomly generated list of kinase–substrates can be found by determining the cardinality of the set of substrates that are targeted by specific kinases (or family of kinases) dividing such number by the total number of substrates in the background dataset. In order to detect statistical significant deviations from this expected value, we use the Fisher Exact Test (Fisher, 1922). The *P*-value can be used to distinguish specific kinases among the large number of kinases appearing in the output table.

To implement the web-based system we use Java Server Pages (JSP) and MySQL database running on a Tomcat server. All reported results can be exported to Excel via CSV files. Additionally, users can mouse over on the number of targets for each kinase, kinase family or class to see the list of substrates and view a connectivity diagram that visualizes known protein–protein interactions within the substrates using a database of protein–protein interactions we previously published (Berger *et al*., 2007). The map is dynamic where users can move nodes around and click on nodes for more detail (Fig. 1). The visualization of these connectivity diagrams was achieved using Adobe Flash CS4 with ActionScript. Such subgraphs can be used to link kinase specific substrates to pathways and complexes.

As prior knowledge is increasingly used to interpret high-throughput results, e.g. Balazsi *et al*. (2008), we anticipate that KEA is going to be especially useful for the analysis of proteomics and phosphoproteomics data. KEA can be used for analyzing multivariate datasets collected on a time-course to observe trends in kinase activity overtime. Results that show changes in kinase enrichment under different conditions can be due to one of the following reasons: change in kinase enzymatic activity, change in kinase subcellular localization or changes in kinase concentration. Furthermore, KEA can help researchers understand how they can perturb cellular systems toward a desired phenotype by targeting a kinase or group of kinases with pharmacological or gene silencing means. Kinase signaling is well-established to be disturbed in many disease states, especially in cancer (Blume-Jensen and Hunter, 2001), while it is apparent that phenotypic integrity is controlled by the activity of the regulated behavior of multiple kinases. Hence, mapping kinase activation patterns based on different experimental conditions and time points when measuring many genes/proteins at once in diseased/perturbed versus normal/control may directly suggest combinations of kinase inhibitors that would shift the cellular state towards a desired phenotype.

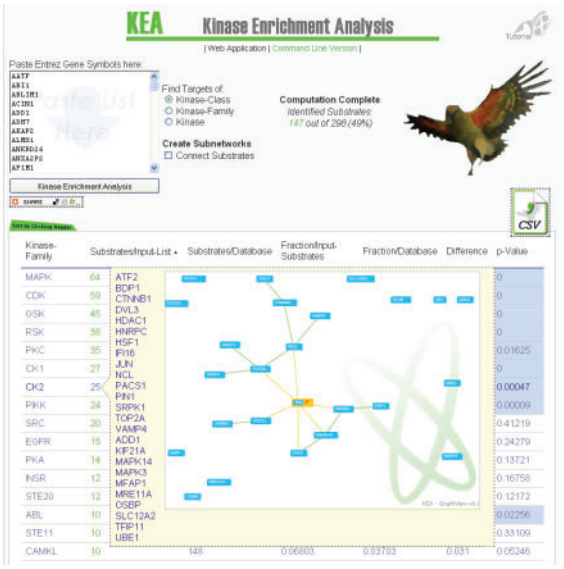

Fig. 1. Screenshot of the KEA user interface. Users can paste lists of Entrez gene symbols, representing human proteins; select the level of analysis: kinase-class, kinase-family or kinase and then the program outputs a list of ranked kinase-classes, kinase-families or kinases based on specificity of phosphorylating substrates from the input list. Substrates can be then connected based on their known protein–protein interaction using an original network viewer developed using Adobe Flash CS4.

ACKNOWLEDGEMENTS

We would like to thank Ben MacArthur, Amin Mazloom, Ihor Lemischka, Kevin Xiao and Robert Lefkowitz for useful discussions.

Funding: National Institutes of Health (Grant No. P50GM071558); Seed fund, Mount Sinai School of Medicine (to A.M.).

Conflict of Interest: none declared.

REFERENCES

Balazsi,G. *et al*. (2008) The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. *Mol. Syst. Biol.*, **4**.

Berger,S. *et al*. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8**, 372.

Blume-Jensen,P. and Hunter,T. (2001) Oncogenic kinase signalling. *Nature*, **411**, 355–365.

Chatr-aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.

Dephoure,N. *et al*. (2008) A quantitative atlas of mitotic phosphorylation. *Proc. Natl Acad. Sci. USA*, **105**, 10762–10767.

Diella,F. *et al*. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.

Fisher,R.A. (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87–94.

Hornbeck,P.V. *et al*. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.

Huang,H.-D. *et al*. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.

Linding,R. *et al*. (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.

Linding,R. *et al*. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.*, **36**, D695–D699.

Ma'ayan,A. *et al*. (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, **309**, 1078–1083.

Manning,G. *et al*. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Mishra,G.R. *et al*. (2008) The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: one small step in manual annotation, one giant leap for full comprehension of genomes. *Mol. Cell Proteomics*, **7**, 1409–1419.

Quintaje,S.B. and Orchard,S. (2008) The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot: One Small Step in Manual Annotation, One Giant Leap for Full Comprehension of Genomes. *Mol. Cell Proteom.*, **7**, 1409–1419.

Yang,C.-Y. *et al*. (2008) PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, **24**, i14–i20.